Evaluating Histopathology Foundation Models for Few-shot Tissue Clustering: an Application to LC25000 Augmented Dataset Cleaning

George Batchkala, Bin Li, Jens Rittscher

IBME/BDI, Department of Engineering Science, University of Oxford, Oxford, UK
{george.batchkala; bin.li; jens.rittscher }@eng.ox.ac.uk

Abstract. Recent digital histopathology datasets have significantly advanced the development of deep learning-based histopathology frameworks. However, data leakage in model training can lead to artificially high metrics that do not genuinely reflect the strength of the approach. The LC25000 dataset, consisting of tissue image tiles extracted from lung and colon samples, is a popular benchmark dataset. In the released version, tissue tiles were augmented randomly and mixed. Nevertheless, many studies report near-perfect accuracy scores, often due to data leakage, where augmented images of the same tissue tile are split into both training and test sets. To improve the quality of performance reports, we develop a semi-automatic pipeline to clean LC25000. By clustering and separating all augmented images of the same tiles, using recently proposed histopathology foundation models and manual correction, we create a clean version of LC25000. We then evaluate the quality of features extracted by these foundational models, using the clustering task as a benchmark. Our contributions are: 1) We publicly release our semi-automatic annotation pipeline along with the LC25000-clean dataset to facilitate appropriate utilization of this dataset, reducing the risk of overestimating models' performance; 2) We profile various combinations of feature extraction and clustering methods for identifying duplicates of the same image generated by basic image transformations; 3) We propose the clustering task as a minimal-setup benchmark to evaluate the quality of tissue image features learned by histopathology foundation models. Clustering labels, annotation pipeline, and evaluation code: https://github.com/GeorgeBatch/LC25000-clean

Keywords: computational pathology · LC25000 · lung · colon · cancer · dataset cleaning · foundation model · representation learning · clustering

1 Introduction

Recent digital histopathology datasets consisting of digitized tissue specimens have facilitated advancements in computational pathology, which is vital for developing and validating deep learning-based frameworks such as tumour detection, cancer subtyping, and therapy response prediction [22]. While these

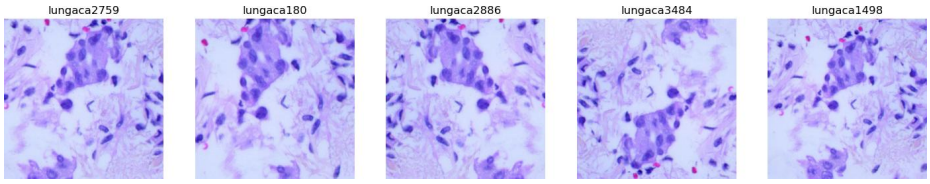

Fig. 1. Examples of augmented images from the same origin tile with their names in the released dataset showing on the top. The images were shuffled and indexed randomly.

datasets enable significant progress, caution is necessary to avoid common pitfalls that may lead to data leakage and invalid model evaluation. The LC25000 dataset [2], for instance, is a widely-used benchmark source, cited over 140 times[1]. It comprises 25,000 tiles extracted from lung and colon histopathology images, divided into five classes with 5,000 images per class: lung adenocarcinoma (lung_aca), lung squamous cell carcinoma (lung_scc), normal lung (lung_n), colon adenocarcinoma (colon_aca), and normal colon (colon_n). This dataset is mainly used in developing tile-level classifiers for cancer tissue classification, with various studies reporting accuracy scores of 95% and above [13,14,16,20,21,24].

However, it is crucial to be aware of issues such as *type-1* data leakage, where the augmented images of the same tissue tile are split into the training and test set, leading to over-estimation of model performance attributing to simple shortcuts that associate these duplicates. Similar concerns apply to datasets like TCGA, where multiple slides from the same patient may inadvertently be included in both training and testing sets (*type-2* data leakage).

According to the original LC25000 publication [2], the authors collected 250 original images for each of the five classes mentioned above, which we will refer to as **prototypes**. They expanded the dataset to 25,000 images through a series of random image transforms, including random rotations and flips. All images were center-cropped and resized to 768x768 pixels. Given that 5,000 images in each class are derived from 250 prototypes, each prototype generates around 20 augmented images. We refer to augmented images of the same tissue tile (a prototype) as **semantic duplicates**. Suppose these 5,000 images are randomly split into training, validation, and test sets using an 80/20 ratio. Then there is a $\sim 99\%$ chance $(1 - p(\text{all 20 in test}) - p(\text{all 20 in train}) = 1 - 0.2^{20} - 0.8^{20})$ semantic duplicates of the same prototype appearing in both the training and test sets. Consequently, we can expect *type-1* data leakage to occur for almost all prototypes when models are both trained and evaluated on LC25000.

Acknowledging the contributions and widespread use of LC25000, we aim to enhance the quality of performance reports when researchers utilize this dataset. Many studies report near-perfect accuracy scores (often 99.9%), primarily due to data leakage. This has unfortunately led to an unnecessary allocation of research

[1] https://www.semanticscholar.org/paper/Lung-and-Colon-Cancer-Histopathological-Image-Borkowski-Bui/6c5a72aaa3c29d52d46ef904f15719478b6cdfc2

resources and reviewing efforts. Therefore, we intend to support the research community in making more accurate and reliable use of the LC25000 dataset.

Inspired by the recent success of automatic data cleaning of the Quilt-1M histopathology dataset [1,9], we develop a semi-automatic pipeline to clean the LC25000 dataset. In our approach, we first cluster the images into groups of images originating from the same prototype and then perform a semi-automatic check and correction, creating LC25000-clean dataset where all semantic duplicates belonging to the same prototypes are grouped together.

Our contributions are: First, we release our semi-automatic annotation pipeline along with **LC25000-clean** dataset to facilitate appropriate utilisation of the LC25000 dataset. Second, we assess various combinations of feature extraction and clustering methods for clustering semantic duplicates. Finally, we propose using the clustering task as a minimal-setup benchmark to evaluate emerging histopathology foundation models. This task can be regarded as a quality lower bound of the tissue image features extracted by these models.

2 Methodology

We describe below our dataset cleaning pipeline, which produces **LC25000-clean**, and evaluation of features extracted by recently proposed histopathology foundation models and general image feature extractor baselines.

2.1 Semi-automatic dataset cleaning

Our first goal is to obtain a clean version of LC25000, where semantic duplicates are grouped according to their prototype memberships.

Feature extraction. The original 768x768 image tiles are resized into 224x224, and image features are computed using pre-trained image models, such as the UNI [3] model. The resulting features are used for clustering to produce an initial cleaning result. We also evaluate the effects of different image normalization options, feature extractors, and dimensionality reduction methods.

Image Pre-processing. We explore three different image RGB normalization options while always resizing the 768x768 images to 224x224: (1) resize-only: no normalization applied after resizing; (2) normalizing with ImageNet constants as suggested by the authors of UNI [3], Prov-GigaPath [23], and Phikon[5] models; (3) normalizing using mean and standard deviation constants computed separately for each of the 5 classes of LC25000.

Feature Extraction. We extract features with models pre-trained on natural images and histopathology-specific models. Natural-image models included ResNet18 [7] and Truncated ResNet50 as used in the CLAM pipeline [12], as well as the small and base versions of ViT pre-trained with DINOv2 [17]. Pathology-specific models included ResNet18 pre-trained with SimCLR [4] on the lung portion of TCGA [11], UNI [3], Prov-GigaPath [23], and Phikon [5].

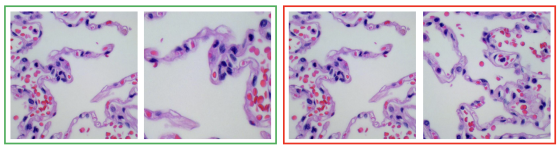

Fig. 2. Normal lung. Left: accepted positive pair. Right: rejected negative pair.

Dimensionality Reduction. The feature extractors described above output feature vectors in different sizes: 512, 768, 1024, 1536. However, the information encoded might be excessive when clustering augmented images. Hence, we test PCA and UMAP for reducing the dimension of extracted features before running the clustering algorithms: PCA with 0.9, 0.95, and 0.99 proportion of variance explained and UMAP [15] with 2, 8, 32 components (other parameters are fixed).

Clustering and manual cleaning. The feature vectors corresponding to the image tiles are fed into K-Means with 250 clusters, producing an initial clustering. The manual stage contains three components: (1) manually accepting or rejecting cluster assignments by iterating through the initial clustering result, (2) automatic clustering of the rejected images and manually purifying rejected clusters, and (3) manually merging pure accepted and rejected clusters.

Manual Accepting and Rejecting of Cluster Assignments. First, we compute the centroid embedding of each cluster c and select the image closest to the centroid as a reference I_r^c. Then, for each cluster c we show a pair $(I_r^c, I_j^c) \forall j \neq r$ and iterate over all other images I_j^c that have been automatically assigned to this cluster. The annotator is asked to choose whether an image comes from the same origin (see Figure 2). If so, image I_j^c is confirmed to belong to cluster c. Otherwise, I_j^c is added to the pool of rejected images. After going through all 250 clusters, we end up with accepted and rejected sets. The accepted set contains 250 pure clusters. The rejected set contains the images incorrectly assigned by the clustering algorithm (2% of image tiles for lung tissue, 3% for colon tissue).

Clustering Rejected Images and Purifying Rejected Clusters. We run the clustering algorithm on the rejected images. Then, an annotator looks through and manually splits impure clusters into pure clusters until only pure clusters remain.

Merging Pure Accepted and Pure Rejected Clusters. Given that the clusters in the accepted and rejected pools are pure, the next step is to merge overgrained clusters (examples of clusters to be merged are shown in Figure 3). We compute pairwise distances between the clusters via single linkage, i.e., assigning the distance between clusters $\mathbf{A}$ and $\mathbf{B}$ with the smallest distance between any two pairs of feature vectors $\mathbf{a_i} \in \mathbf{A}$ and $\mathbf{b_j} \in \mathbf{B}$, and check the closest pairs of clusters to decide if a paired clusters should merge.

To achieve the least manual effort, our final pipeline consists of feeding resized images directly into the UNI feature extractor, reducing the output features using UMAP with 8 components, and running K-Means on the reduced features.

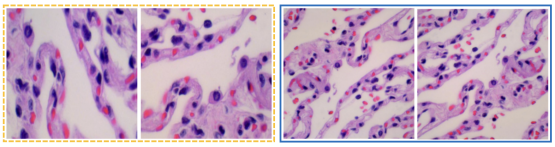

Fig. 3. Two pure normal lung tissue clusters that needed merging.

2.2 Evaluating features extracted by different foundation models

Clustering semantic duplicates according to prototypes. We profile the quality of image features learned by recent histopathology foundation models clustering LC25000 and testing the clustering results against LC25000-clean. We also test the effects of different image normalization methods. In a perfect scenario, LC25000 can be automatically (without manual correction) clustered into LC25000-clean, where each cluster contains all semantic duplicates generated by random image transformation from the same prototype.

For the clustering assignment A_p, we compute a connectivity matrix between all pairs of images where any two images belonging to the same cluster are connected. We then evaluate this assignment against the ground truth assignment A_m. Binary connectivity (Accuracy, Precision, Recall, F1-score, Specificity, Balanced Accuracy) and clustering (Fowlkes–Mallows index [6], Adjusted Rand Index [8,18], Normalized Mutual Information score, Homogeneity, Completeness, V-Measure [19]) are used to evaluate the clustering performance. We further incorporate precision@1 and precision@5 for evaluation, as they are agnostic to the clustering algorithm (e.g., precision@1 only checks the fraction of samples whose nearest neighbour's label is the same as the sample itself). Fowlkes–Mallows Index (**FMI**) is used as the main metric for assessing clustering performance, as it is a balanced metric and reflects the global clustering quality. To compute FMI, the best alignment of the two clustering assignments is found using the Hungarian Algorithm [10]. Then, the numbers of true positives (pairs of points that are in the same cluster in both A_m and A_p), false positives (pairs of points that are in the same cluster in A_p, but in different clusters in A_m), and false negatives (pairs of points that are in different clusters in A_p, but in the same cluster in A_m) are calculated. Finally, FMI is calculated as $FMI = \sqrt{(TP/(TP+FP)) \times (TP/(TP+FN))} = \sqrt{Precision \times Recall}$.

Tumor classification on the LC25000-clean dataset. LC25000 was originally proposed for two classification tasks: three-class lung tissue classification (adenocarcinoma, squamous cell carcinoma, normal lung tissue) and binary colon tissue classification (adenocarcinoma vs normal colon tissue). To evaluate the effect of *type-1* data leakage on the reporting performance of the classification model on the original LC25000 dataset, we train classifiers using KNN and Linear probing based on raw image features extracted by the foundation models.

Randomly splitting the original LC25000 dataset images into train and test sets results in both sets containing semantic duplicates from the same prototypes.

In this circumstance, KNN with 1 nearest neighbour can achieve an almost perfect accuracy score, because, for each image, the nearest neighbour is always one of its semantic duplicates (thus, having the same class label). For linear probing, we train a single linear layer with softmax activation for the multi-class lung tissue classification and sigmoid activation for the binary colon tissue classification. Similar to the case of KNN, the linear layer should be able to overfit the training data, thereby overfitting the test data provided that most samples in the test set will have semantic duplicates in the training set. Therefore, we overfit by training the model until the training loss converges (no validation).

In contrast, on the clean dataset, it will be harder for the KNN classifier with 1 nearest neighbour to achieve perfect accuracy because no image in the test set will be a semantic duplicate of any images in the training set. For the linear classifier, overfitting the training set is expected to result in a performance drop.

To clearly delineate how much of the previously performance reported performance is attributed to the data leakage, we add further stress tests by decreasing the ratio of training samples, starting with a popular 80/20 train/test split and decreasing the ratio to 5/95, where only 5% of samples are used for training.

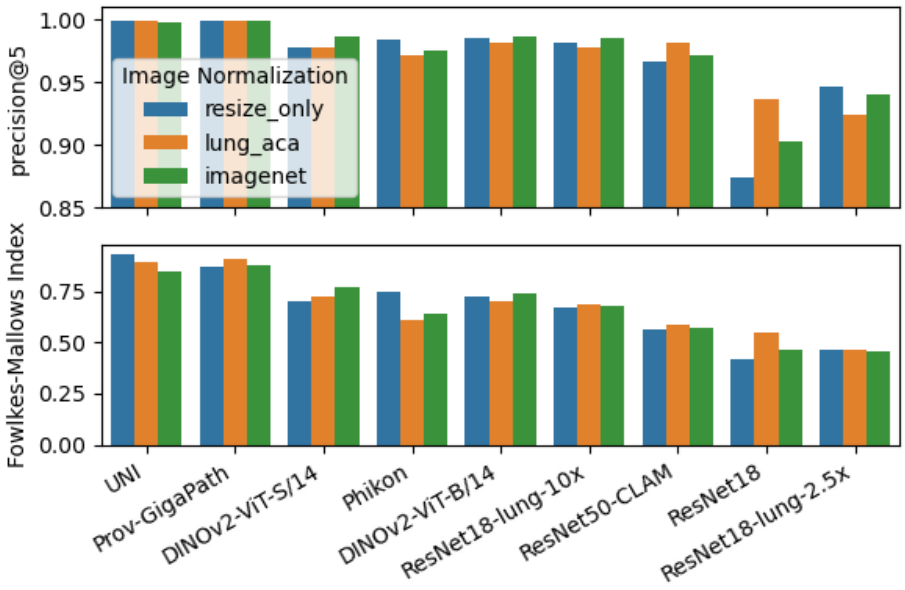

Fig. 4. Precision@5. Outputs from feature extractors are ranked by how close they are in Euclidean distance. **Fawlkes-Mallows Index.** Raw outputs from feature extractors are passed into K-Means clustering (Euclidean distance). **Image Normalization methods.** *resize_ only*: using raw RGB values (0 to 1). *lung_ aca*: using the statistics computed from the LC25000 dataset (mean and variance) to center the inputs and make a unit variance. *imagenet*: using the statistics of the ImageNet dataset.

3 Results and Discussions

3.1 Clustering semantic duplicates into prototypes.

After obtaining LC25000-clean where all images are grouped and assigned prototype labels, we evaluate the quality of features extracted using different foundation models by clustering the original LC25000 dataset and compare the clustering results against LC25000-clean.

The raw features extracted by each model are clustered using K-Means algorithm. We noticed that the results vary for the same feature extractor depending on the image normalization method, with FM-index reported in Figure 4. Figure 5 shows the clustering performance under different metrics, with each feature extractor's best-performing image normalization technique (by FM-index).

UNI and Prov-GigaPath outperform other models by large margins. Interestingly, the best image normalization method for UNI is no RGB normalization, whereas for Prov-GigaPath it is to normalize the input using the corresponding dataset's statistics such that the dataset is centered and has a unit variance (i.e., the mean and variance of the LC25000 lung dataset). Note that both models were trained with inputs normalized using ImageNet statistics [3,23].

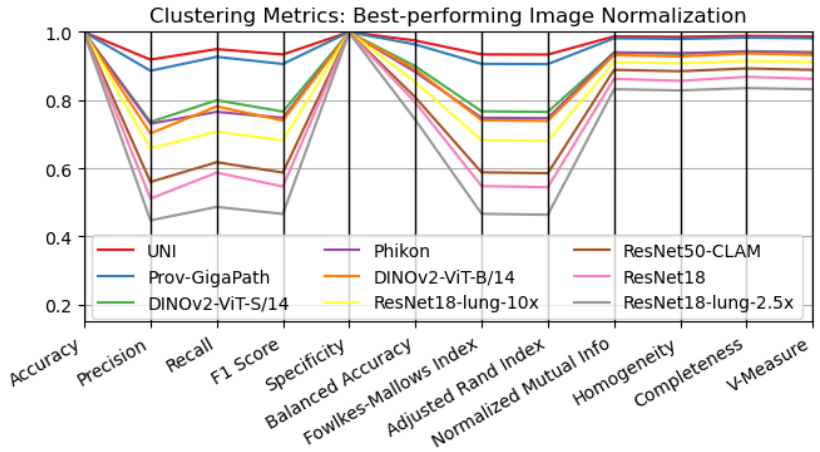

Fig. 5. Clustering performance: all metrics. Outputs from feature extractors are passed directly into K-Means clustering. For each feature extractor, the results correspond to using the best normalization-extractor combination (by FM-index - see Figure 4).

3.2 KNN-1 and Linear Probing

Compared to ResNet18, the lung tissue classes are much better separated in feature space by Phikon and UNI, showing better feature qualities from pathology-specific feature extractors as shown in Figure 6.

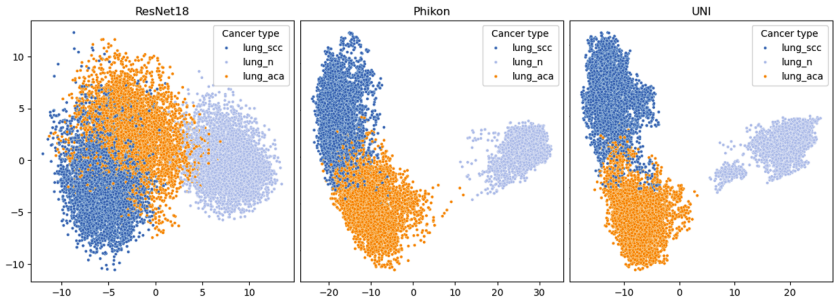

Fig. 6. First 2 PCA projections of the raw feature outputs of ResNet18, Phikon, and UNI feature extractors (images were normalized using ImageNet constants).

Table 1 shows the classification accuracy achieved by KNN (1 nearest neighbour) and linear classifiers on the features computed from ImageNet pre-trained ResNet18, Phikon, and UNI feature extractors on original and cleaned versions of the LC25000 dataset (images were normalized using ImageNet constants, same for all models to facilitate the comparison). These 3 models are chosen based on the clustering performance of the lung adenocarcinoma class, and on that, they have different representation strengths: ResNet18 (weak), Phikon (medium), and UNI (strong), as indicated in Figure 4. We vary the train/test split proportions to show how easy/difficult it is to achieve near-perfect prediction accuracy on the original and cleaned versions of the dataset. The reported values are computed from 10 random draws for each split proportion, with the mean and standard deviation of the accuracy values.

			Lung		**Colon**	
Split	**Features**	**CLS**	**Original**	**Clean**	**Original**	**Clean**
80% - 20%	ResNet18	KNN	0.998 ± 0.001	0.917 ± 0.007	0.999 ± 0.000	0.962 ± 0.007
		Linear	0.970 ± 0.003	0.943 ± 0.007	0.998 ± 0.001	0.993 ± 0.003
	Phikon	KNN	**1.0 ± 0.0**	0.993 ± 0.003	**1.0 ± 0.0**	**1.0 ± 0.0**
		Linear	**1.0 ± 0.0**	0.987 ± 0.009	**1.0 ± 0.0**	**1.0 ± 0.0**
	UNI	KNN	**1.0 ± 0.0**	0.994 ± 0.005	**1.0 ± 0.0**	**1.0 ± 0.0**
		Linear	**1.0 ± 0.0**	0.989 ± 0.01	**1.0 ± 0.0**	**1.0 ± 0.0**
20% - 80%	ResNet18	KNN	0.981 ± 0.002	0.892 ± 0.009	0.992 ± 0.001	0.935 ± 0.013
		Linear	0.961 ± 0.002	0.923 ± 0.007	0.995 ± 0.001	0.981 ± 0.002
	Phikon	KNN	**1.0 ± 0.0**	0.972 ± 0.006	**1.0 ± 0.0**	0.999 ± 0.001
		Linear	0.998 ± 0.000	0.977 ± 0.004	**1.0 ± 0.0**	0.999 ± 0.001
	UNI	KNN	**1.0 ± 0.0**	0.978 ± 0.005	**1.0 ± 0.0**	**1.0 ± 0.0**
		Linear	0.999 ± 0.000	**0.979 ± 0.003**	**1.0 ± 0.0**	0.999 ± 0.001
5% - 95%	ResNet18	KNN	0.940 ± 0.005	0.862 ± 0.020	0.970 ± 0.002	0.872 ± 0.038
		Linear	0.945 ± 0.002	0.894 ± 0.019	0.984 ± 0.002	0.942 ± 0.020
	Phikon	KNN	0.994 ± 0.001	0.948 ± 0.013	**1.0 ± 0.0**	0.994 ± 0.004
		Linear	0.993 ± 0.002	0.960 ± 0.005	**1.0 ± 0.0**	0.992 ± 0.005
	UNI	KNN	**0.996 ± 0.001**	**0.962 ± 0.008**	**1.0 ± 0.0**	**0.997 ± 0.002**
		Linear	0.994 ± 0.002	0.960 ± 0.014	**1.0 ± 0.0**	0.990 ± 0.008

Table 1. Classification accuracy (mean ± st.d.) computed on 10 random splits of ResNet18, Phikon, and UNI extractors with ImageNet image normalization.

We observe that the classification accuracy ordering of the feature extractors is consistent with the clustering performance: UNI performs best, Phikon follows, while ResNet18 pre-trained on ImageNet shows the worst results. For the original dataset, the models all can achieve very high accuracy; even with only 5% of the images for training, a simple KNN classifier can still achieve a near-perfect performance (>99.5% accuracy), simply due to overfitting and data leakage. As expected, the classification accuracy drops significantly on the cleaned dataset compared to the original dataset which is subjected to *type-1* data leakage, highlighting the importance of appropriate treatment of the dataset for a robust training/validation split that reflects the true power and generalizability of the evaluated models.

Acknowledgments. George Batchkala is supported by Fergus Gleeson's A2 research funds, UKRI DART Lung Health Program (Innovate UK grant 40255), and the EPSRC Center for Doctoral Training in Health Data Science (EP/S02428X/1).

Disclosure of Interests. There are no conflicts of interest.

References

1. Aubreville, M., Ganz, J., Ammeling, J., Kaltenecker, C., Bertram, C.: Model-based Cleaning of the QUILT-1M Pathology Dataset for Text-Conditional Image Synthesis. In: Medical Imaging with Deep Learning (Apr 2024)
2. Borkowski, A.A., Bui, M.M., Thomas, L.B., Wilson, C.P., DeLand, L.A., Mastorides, S.M.: Lung and Colon Cancer Histopathological Image Dataset (LC25000). arXiv:1912.12142 [cs, eess, q-bio] (Dec 2019)
3. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F.K., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., Williams, M., Oldenburg, L., Weishaupt, L.L., Wang, J.J., Vaidya, A., Le, L.P., Gerber, G., Sahai, S., Williams, W., Mahmood, F.: Towards a general-purpose foundation model for computational pathology. Nature Medicine **30**(3), 850–862 (Mar 2024). https://doi.org/10.1038/s41591-024-02857-3
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A Simple Framework for Contrastive Learning of Visual Representations. In: Proceedings of the 37th International Conference on Machine Learning. pp. 1597–1607. PMLR (Nov 2020)
5. Filiot, A., Ghermi, R., Olivier, A., Jacob, P., Fidon, L., Kain, A.M., Saillard, C., Schiratti, J.B.: Scaling Self-Supervised Learning for Histopathology with Masked Image Modeling (Sep 2023). https://doi.org/10.1101/2023.07.21.23292757
6. Fowlkes, E.B., Mallows, C.L.: A Method for Comparing Two Hierarchical Clusterings. Journal of the American Statistical Association **78**(383), 553–569 (Sep 1983). https://doi.org/10.1080/01621459.1983.10478008
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
8. Hubert, L., Arabie, P.: Comparing partitions. Journal of Classification **2**(1), 193–218 (Dec 1985). https://doi.org/10.1007/BF01908075
9. Ikezogwo, W.O., Seyfioglu, M.S., Ghezloo, F., Geva, D.S.C., Mohammed, F.S., Anand, P.K., Krishna, R., Shapiro, L.: Quilt-1M: One Million Image-Text Pairs for Histopathology. In: Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Nov 2023)
10. Kuhn, H.W.: The Hungarian method for the assignment problem. Naval Research Logistics Quarterly **2**(1-2), 83–97 (Mar 1955). https://doi.org/10.1002/nav.3800020109
11. Li, B., Li, Y., Eliceiri, K.W.: Dual-Stream Multiple Instance Learning Network for Whole Slide Image Classification With Self-Supervised Contrastive Learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14318–14328 (2021)
12. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature Biomedical Engineering 2021 5:6 pp. 555–570 (Mar 2021). https://doi.org/10.1038/s41551-020-00682-w
13. Mangal, S., Chaurasia, A., Khajanchi, A.: Convolution Neural Networks for diagnosing colon and lung cancer histopathological images. ArXiv (Sep 2020)
14. Masud, M., Sikder, N., Nahid, A.A., Bairagi, A.K., AlZain, M.A.: A Machine Learning Approach to Diagnosing Lung and Colon Cancer Using a Deep Learning-Based Classification Framework. Sensors **21**(3), 748 (Jan 2021). https://doi.org/10.3390/s21030748

15. McInnes, L., Healy, J., Melville, J.: UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction (Sep 2020). https://doi.org/10.48550/arXiv.1802.03426
16. Mehmood, S., Ghazal, T.M., Khan, M.A., Zubair, M., Naseem, M.T., Faiz, T., Ahmad, M.: Malignancy Detection in Lung and Colon Histopathology Images Using Transfer Learning With Class Selective Image Processing. IEEE Access **10**, 25657–25668 (2022). https://doi.org/10.1109/ACCESS.2022.3150924
17. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning Robust Visual Features without Supervision (Feb 2024). https://doi.org/10.48550/arXiv.2304.07193
18. Rand, W.M.: Objective Criteria for the Evaluation of Clustering Methods. Journal of the American Statistical Association **66**(336), 846–850 (Dec 1971). https://doi.org/10.1080/01621459.1971.10482356
19. Rosenberg, A., Hirschberg, J.: V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure. In: Conference on Empirical Methods in Natural Language Processing (Jun 2007)
20. Talukder, M.A., Islam, M.M., Uddin, M.A., Akhter, A., Hasan, K.F., Moni, M.A.: Machine learning-based lung and colon cancer detection using deep feature extraction and ensemble learning. Expert Systems with Applications **205**, 117695 (Nov 2022). https://doi.org/10.1016/j.eswa.2022.117695
21. Toğaçar, M.: Disease type detection in lung and colon cancer images using the complement approach of inefficient sets. Computers in Biology and Medicine **137**, 104827 (Oct 2021). https://doi.org/10.1016/j.compbiomed.2021.104827
22. van der Laak, J., Litjens, G., Ciompi, F.: Deep learning in histopathology: The path to the clinic. Nature Medicine **27**(5), 775–784 (May 2021). https://doi.org/10.1038/s41591-021-01343-4
23. Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y., Xu, Y., Wei, M., Wang, W., Ma, S., Wei, F., Yang, J., Li, C., Gao, J., Rosemon, J., Bower, T., Lee, S., Weerasinghe, R., Wright, B.J., Robicsek, A., Piening, B., Bifulco, C., Wang, S., Poon, H.: A whole-slide foundation model for digital pathology from real-world data. Nature **630**(8015), 181–188 (Jun 2024). https://doi.org/10.1038/s41586-024-07441-w
24. Yu, G., Sun, K., Xu, C., Shi, X.H., Wu, C., Xie, T., Meng, R.Q., Meng, X.H., Wang, K.S., Xiao, H.M., Deng, H.W.: Accurate recognition of colorectal cancer with semi-supervised deep learning on pathological images. Nature Communications **12**(1), 6311 (Nov 2021). https://doi.org/10.1038/s41467-021-26643-8